# 科学数据开放共享中的数据质量治理研究*

■ 盛小平[1] 田婧[1] 向桂林[2]

[1] 上海大学图书情报档案系 上海 200444 [2] 中国科学院生物物理研究所 北京 100101

摘 要：[目的/意义]探究科学数据开放共享中的数据质量问题及其治理对策，以便促进科学数据开放共享的有效实施。[方法/过程]运用规范分析法和因果分析法，分析当前科学数据开放共享中的数据质量问题和引发问题的根本原因，构建科学数据开放共享数据质量治理模型，并从诱因入手提出4类治理对策。[结果/结论]科学数据开放共享中的数据质量问题涉及科学数据的准确性、完整性、一致性、及时性、可靠性、关联性、开放可访问性。可以从政策法规、组织管理、技术与平台、利益相关者4个方面制定科学数据质量治理对策，从而解决相关科学数据质量问题，进一步推动科学数据开放共享的实施。

关键词：科学数据 开放共享 数据质量 质量治理 治理对策

分类号：G203

DOI：10.13266/j.issn.0252-3116.2020.22.002



开放数据作为数字时代的社会资本，已经成为推动社会经济发展的重要因素[1]。而数据质量成为影响开放数据及其共享效果的关键，越来越受到人们的高度重视。国际标准化组织（International Organization for Standardization，简称ISO）制定了ISO 8000数据质量标准，致力于从数据质量活动、数据质量原则、数据质量特征等角度管理数据质量[2]。《G8开放数据宪章》宣告将及时、全面、准确地发布高质量的开放数据，满足最高标准的开放数据质量要求[3]。2018年我国实施的《科学数据管理办法》明确要求“按照有关标准规范进行科学数据采集生产、加工整理和长期保存，确保数据质量”“法人单位应建立科学数据质量控制体系，保证数据的准确性和可用性”[4]。事实上，高质量的科学数据既是科学研究的基础，也是科学研究的驱动力[5]，还是成功的基本要素[6]。然而，数据质量障碍已成为目前科学数据开放共享中的一个主要问题[7-8]。依据数据仓储研究所（The Data Warehousing Institute）的观点，数据质量问题的成本每年超过6 000亿美元[9]。虽然已有一些文献探讨了数据质量管理问题，但鲜见论述科学数据开放共享中的数据质量治理。因此，针对目前科学数据开放共享中的数据质量问题，应该加强科学数据质量治理研究，以便进一步推动科学数据开放共享、开放研究与开放创新。

## 1 科学数据开放共享中的数据质量问题及其缘由

了解科学数据开放共享中的数据质量问题及其缘由是实施科学数据质量治理的前提。

### 1.1 科学数据开放共享中的数据质量问题

数据质量是指数据在使用过程中满足特定目的的需求的程度[10]。数据质量维度衡量数据在某一方面的性质，可以为数据质量的业务需求提供框架，便于对质量进行量化度量[11]。数据质量属性有多种多样，包括可访问性、准确性、现实性、可用性、可信性、明确性、完整性、综合性、一致性、正确性、及时性、易用性、灵活性、互用性、可解释性、易学性、精确性、不重复、客观性、冗余度、关联性、安全性、时事性、可追踪性、效用性、有效性、价值性等[12]。《国际标准化组织/国际电工委员会（ISO/IEC）25012》标准将数据质量属性分为三类[13]：①内在数据质量属性，包括准确性、完整性、一致性、可信性、即时性；②系统相关数据质量属性，包括有效性、可携带性、可恢复性；③内在的与系统相关的数据质量属性，包括可访问性、兼容性、机密性、效率

* 本文系国家社会科学基金项目“开放科学环境下的科学数据开放共享机制与对策研究”（项目编号：18ATQ007）研究成果之一。

作者简介：盛小平（ORCID：0000-0002-6341-6973），教授，博士，博士生导师，E-mail：shengxp68@126.com；田婧（ORCID：0000-0002-3760-5308），硕士研究生；向桂林（ORCID：0000-0002-0880-8106），副研究馆员，博士。

收稿日期：2020-06-09 修回日期：2020-07-21 本文起止页码：11-24 本文责任编辑：易飞

性、精确性、可追溯性、可理解性。任何组织都必须重视数据质量,这是因为高质量数据既是有价值的资产,可以成为战略性的竞争优势,也能提高客户满意度,还能增加收入和利润[14]。与此相反,未经识别和纠正的糟糕质量的数据可能对组织产生重大的负作用,比如,降低客户满意度,降低决策过程效率,降低绩效,降低雇员工作满意度,增加运营成本,对数据失去信任,对组织文化产生负面影响等[15]。

国外学者认为,为了满足预期的使用,数据必须具有准确性、及时性、相关性、完整性、可信性和可理解性[16]。一般科学数据质量问题通常涉及数据的准确性、完整性、一致性、及时性、可靠性、关联性等[17]。其中,数据准确性是指数据正确、可靠、无误;数据完整性是指数据有足够的深度、广度和范围来完成当前任务;数据一致性是指数据总是以相同的格式显示,并且与以前的数据兼容;数据及时性是指数据的新旧适合于当前任务;数据可靠性是指数据能够传达正确的信息,是可信赖的或可信任的;数据关联性是指数据对当前任务是适用的和有用的[18]。

上述数据质量问题同样可出现在科学数据开放共享活动中,因为开放科学数据必须保障数据的准确性、完整性、一致性、及时性、可靠性、关联性。除此之外,开放科学数据质量问题还涉及数据的开放可访问性[19]。它是指数据是容易利用和快速检索,并且能够以公开、免费或开放获取的方式得到发布与传播。按照上述 7 类数据属性,可以进一步描述科学数据开放共享可能遇到的各种数据质量问题如表 1 所示:

表 1 科学数据开放共享中的数据质量问题一览

| 数据属性问题分类 | 科学数据开放共享中的数据质量问题描述 | 文献来源 | 数据属性问题分类 | 科学数据开放共享中的数据质量问题描述 | 文献来源 |
|---|---|---|---|---|---|
| 准确性问题 | 数据模糊或错误 | [19 – 21] | 开放可访问性问题 | 缺少机器可读格式 | [19,25] |
| | 字序或内容错误 | [19,22 – 23] | | 数据无法下载 | [19] |
| | 数据污染 | [24] | | 有限的检索能力 | [41] |
| | 数据或数据集碎片化 | [8,25 – 26] | | 缺少开放共享平台 | [8,39] |
| | 编码不准确 | [19,21,27] | | 数据软件不兼容 | [8,39] |
| | 数据输入错误 | [19,27] | | 数据开放程度较低 | [42] |
| | 数据更新/传输/分类错误 | [20] | | 不能访问原始数据,只能访问处理后的数据 | [8] |
| | 三元组抽取错误 | [28] | | 获取开放知识库存储的原始数据困难 | [43] |
| | 数据类型抽取不正确 | [28] | | 缺少开放数据质量标准 | [44] |
| | 语法错误 | [29 – 30] | | | |
| | 冗余实例 | [29] | | | |
| | 引用不正确 | [27,31] | | | |
| 一致性问题 | 数据不一致 | [22,32] | 完整性问题 | 数据不完整 | [8,22,45,39] |
| | 数据结构不一致 | [27,33] | | 缺少数据质量信息 | [25] |
| | 数据格式不一致 | [8,19,21-22,25,27,33 – 34] | | 缺少元数据 | [7,25,39,46 – 47] |
| | 数据标准不一致 | [33 – 34] | | 缺少数据索引 | [39] |
| | 抽取出的信息不相关 | [28] | | 缺少数据值 | [19,22] |
| | 虚假或错误的注释 | [35] | | | |
| | 数据重复 | [22,36 – 37] | | | |
| 可靠性问题 | 无用或无关的数据 | [20,31] | 关联性问题 | 缺少数据链接 | [28,37] |
| | 缺少许可 | [28] | | 数据关联错误 | [7,28,37] |
| | 不明晰的所有权 | [38] | | 死链接、断开的链接和不可引用的坏链接 | [48] |
| | 不正确或不完整的属性值 | [29] | | 存在大量的信息孤岛 | [36] |
| | 未知的数据位置 | [38] | | | |
| 及时性问题 | 数据及时性差 | [19,25,39] | | | |
| | 数据时效性变短 | [40] | | | |

盛小平，田婧，向桂林.科学数据开放共享中的数据质量治理研究[J].图书情报工作,2020,64(22):11－24.

### 1.2 科学数据开放共享中数据质量问题的缘由分析

通常情况下，手工数据输入、初始数据转换、系统整合、数据处理、数据清理、数据净化、系统更新、新数据使用、流程自动化等环节或因素都可影响数据质量[49]。在科学数据开放共享过程中，引发数据质量问题的根本原因主要包括如下4个方面：

#### 1.2.1 政策法规因素

有效的科学数据管理政策与法规是高质量科学数据的根本保障，这是因为它们对开放科学数据管理具有规范与指导作用，而且可以明确利益相关者在科学数据管理中的权利与义务以及风险和合规问题[50]。虽然目前我国已经颁布了《科学数据管理办法》，但是并没有建立科学数据质量控制体系来对数据质量进行度量，也没有建立健全科学数据或个人数据保护法律体系来为创建、共享与利用高质量的科学数据提供法律保障。《中国科学院科学数据管理与开放共享办法》可以用来指导与规范中国科学院系统内成员单位与个人的科学数据管理与开放共享行为，但很少有其他机构制定类似的科学数据管理实施细则来提升科学数据质量管理水平。现行的《政府信息公开条例》没有有效区分与界定开放政府数据中的科学数据类型及其质量要求，势必会对开放政府科学数据质量造成影响。

#### 1.2.2 组织管理因素

许多人认为大多数数据质量问题是由数据输入错误引起的，然而，实际上许多数据质量问题是由于缺乏对高质量数据的组织承诺造成的，而后者本身源于治理和管理领导力的缺乏[27]，包括缺乏开放科学数据质量管理的统一领导与部门协调，缺少有效的开放科学数据质量管理制度，没有建立科学数据质量管理计划与流程，缺少开放科学数据质量标准，缺乏有效的激励机制，在数据生产者和管理者之间缺少互惠机制，没有建立数据质量治理的组织、制度、标准和技术手段[40]，以及陈旧的业务规则，执行不一致的业务流程，缺乏数据操作流程知识的培训[27]，缺少质量管理的组织文化[51]等，所有这些都有可能引起开放科学数据质量问题。

#### 1.2.3 技术与平台因素

高质量的开放科学数据既依赖于组织文化，也依赖于信息技术与共享平台的支撑与有效利用，特别是在科学数据提交、存储、分享、使用和维护等环节上。然而，在目前科学数据开放共享过程中，常常存在诸多技术与平台问题，比如：数据输入接口没有编辑或控制功能来防止不正确的数据被存入系统中，数据接口没有升级以适应新业务流程变化的需求，为不同的业务目的重复使用字段而不是改变数据模型和用户界面或代码，源系统可能在没有告知下游消费者或说明变更情况下更改数据结构，未能执行引用完整性或关闭验证，未能对实例唯一性进行足够的检查或者关闭了数据库中的唯一约束，编码不准确和空白，数据模型不准确，时间数据不匹配，数据修复中故意输入错误数据或安全漏洞[27]，提供的数据格式不标准，新旧数据格式不兼容，缺乏元数据，缺乏数据标准化技术等[33]。这些因素都可能直接导致科学数据开放共享中的数据质量问题。

#### 1.2.4 利益相关者因素

不同的利益相关者，如政府、研究人员、研究机构、数据中心、图书情报机构、资助机构、出版社、数据专业人员等，分别在数据质量管理过程中扮演不同角色、发挥不同作用。

(1)政府。政府通过颁布相关法律法规、制定科学数据质量标准或管理条例，可以在全国或本地区确保科学数据质量。换句话说，若政府在科学数据质量管理方面不作为，那么不可能在全国或本地区营造良好的科学数据质量政策环境，更不能阻止各种损害科学数据质量行为的发生。

(2)研究人员和研究机构。研究人员和研究机构如何看待科学数据质量问题，是否拥有科学数据质量意识，是否制定科学数据质量管理政策、计划或策略，机构科学数据质量管理政策是否明确了利益相关者的职责和义务或任务，是否在实践中严格遵循科学数据质量标准，是否将开放获取高质量的科学数据作为职业发展的正式标准，都将直接影响研究人员或研究机构创建的科学数据质量，同时也将制约他们能否为科学数据的高效利用提供质量保障。

(3)数据中心和图书情报机构。数据中心与图书情报机构是否建立了科学数据质量保证机制，是否建立了高质量的开放科学数据知识库，是否能够为用户提供一系列高质量的科学数据，是否为用户提供科学数据质量管理培训或科学数据质量评价服务以帮助用户提高其数据质量，这些因素都将对科学数据质量产生直接或间接的影响。

(4)资助机构。资助机构是否制定优先资助政策来保障高质量科学数据的开放获取与保存，是否从引领开放科学范式的角度提高人们的数据质量意识并提高开放科学数据质量，是否积极促进利益相关者在保

证科学数据质量方面的合作,都会对科学数据质量产生积极或消极作用。

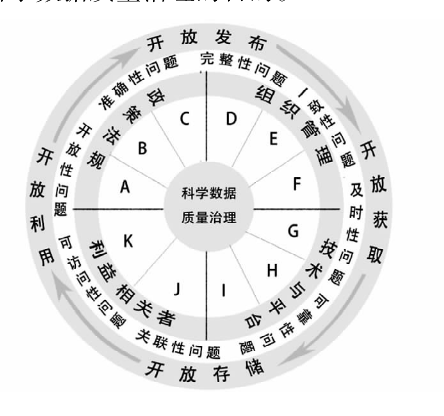
(5)出版社。出版社是否制定支持高质量科学数据开放出版的强制性政策,是否提供高质量的开放获取期刊来改善学术交流基础设施,是否与经过认证的存储库和数据中心协作以简化数据提交,是否通过建立同行评审制度来支持科学数据作为一流的学术产出,是否制定需要关联开放引用科学数据的政策,是否制定鼓励使用文本与数据挖掘的许可政策,都将正向或反向影响科学数据质量[52]。

(6)数据专业人员。数据专业人员的成功参与在很大程度上决定了科学数据质量的命运。他们凭借其专业技能与创造性和系统性工作,如模型构建、规划、问题解决、快速学习、团队合作、适应性和灵活性、注意细节、研究与报告、掌握多种形式的大数据、熟悉尚未解决的问题等[53],可以在科学数据质量管理生命周期中发挥独特的作用,从而有效提升科学数据质量或为科学数据提供质量保证。

## 2 科学数据开放共享中的数据质量治理模型的构建

确保高质量数据是一个复杂的过程。数据治理是任何确保数据质量工作的必要组成部分,具有改进数据质量的潜力,在保障数据质量方面起着非常重要的作用,主要包括:①定义与组织(如企业)数据的使用和管理相关的决策权力和职责[54],有助于更好地实施决策和保护利益相关者需求;②能为组织范围内或全球范围内的数据处理制定和实施数据质量管理规程、指南和路线图;③迫使组织建设数据质量文化,促使人们更广泛地思考质量和重新检查他们的日常实践[55];④对组织中数据的可用性、相关性、使用性、完整性和安全性等进行全面管理,使数据资产价值最大化[56];⑤保证数据是可信的,并确保低质量数据当事人应承担相应的责任[57]。因此,数据治理是解决上述开放科学数据质量问题的重要手段。

如何利用数据治理来确保科学数据开放共享中的数据质量?这需要构建一种科学、实用的数据质量治理模型。这里以科学数据开放共享中的数据质量问题为导向,以科学数据质量治理对策为抓手,建立如图 1 所示的科学数据开放共享中的数据质量治理模型。这种治理模型的核心内容要素包括如下 3 个方面:①科学数据开放共享活动。科学数据开放共享是与科学数据开放生产、组织、发布(或出版)、传播、利用直接相关的一系列价值创造活动,主要包括科学数据的开放发布、开放获取、开放存储、开放利用。而这些科学数据开放共享活动都与科学数据质量有关,需要采取不同的数据治理措施来提升数据质量。②科学数据质量问题。模型以科学数据开放共享中的数据准确性、完整性、一致性、及时性、可靠性、关联性、开放可访问性作为衡量数据质量的维度,并以这些维度的数据质量问题作为数据质量治理的主要目标。③科学数据质量治理措施。针对科学数据开放共享中的主要数据质量问题,结合产生数据质量问题的 4 类缘由,分别从政策法规、组织管理、技术与平台、利益相关者 4 个方面拟定科学数据开放共享中的数据质量治理措施,最终达到实现科学数据质量治理的目的。




注:A:制定《数据质量法》或《数据质量管理条例》;B:制定《开放政府数据法》等相关法律;C:修订与完善《科学数据管理办法》等法规;D:制定科学数据质量战略,明确科学数据质量治理重点与方向;E:建立健全科学数据质量治理结构,明确治理主体的职责与作用;F:制定科学数据质量治理计划,明确科学数据质量治理路径;G:把数据剖析嵌入科学数据质量管理流程;H:实施科学数据质量审计,明确数据质量治理的重点领域;I:构建关联开放数据;J:明确不同利益相关者的科学数据质量治理职责与作用;K:建立利益相关者科学数据质量协同治理机制

图 1 科学数据开放共享中的数据质量治理模型

上述科学数据质量治理模型的运行机制及其主要特点包括:①以科学数据质量治理为"轴",以科学数据质量治理措施为"辐",以各种科学数据属性问题为"内辋",并以科学数据开放共享活动(主要是开放发布、开放获取、开放存储、开放利用)为"外辋",形成科学数据开放共享中的数据质量治理车轮模型。②通过数据质量治理车轮模型及其车轮结构关系,使各种科学数据质量问题与科学数据质量治理措施融合起来。例如,针对科学数据开放共享中的数据质量问题,通过从政策法规上采用科学数据质量治理的对策,比如"制

盛小平，田婧，向桂林. 科学数据开放共享中的数据质量治理研究[J]. 图书情报工作，2020，64(22)：11－24.

定《数据质量法》或《数据质量管理条例》"，可以为解决科学数据准确性、完整性、一致性、及时性、可靠性、关联性、开放可访问性等各种问题提供专门的法律支撑。③通过数据质量治理车轮模型及其车轮结构关系，使科学数据开放共享活动与科学数据质量问题及其治理措施融合起来。例如，作为一种科学数据开放共享方式，科学数据开放出版会产生包括数据准确性、完整性、一致性、及时性、可靠性、关联性、开放可访问性等在内的多种数据质量问题。要解决这些问题，需要从多方面(即产生数据质量问题的 4 类缘由)而非单方面采取有效的数据治理措施。这也适用于其他科学数据开放共享活动。④科学数据开放共享中的数据质量治理是动态变化与循环发展的。正如车轮循环运转方式一样，科学数据开放共享中的数据质量问题与数据质量治理对策是动态变化的、循环发展的。

## 3 科学数据开放共享中的数据质量治理对策

基于上述科学数据质量治理模型，可以从政策法规、组织管理、技术与平台、利益相关者 4 个方面来实施科学数据质量治理。

### 3.1 政策法规方面的治理对策

与数据质量管理相关的法律、法规、规章可以从宏观上给科学数据开放共享中的数据质量治理提供法律保障。这方面的治理对策包括：

#### 3.1.1 制定《数据质量法》或《数据质量管理条例》，为科学数据质量治理提供专门的法律支撑

为确保美国联邦机构使用和传播准确的信息，2000 年底，美国国会批准了"数据质量法"(Data Quality Act，DQA)。DQA 要求联邦机构发布信息质量指南，确保他们传播的信息的质量、实用性、客观性和完整性，并为受影响的人提供纠正这些信息的机制[58]。不过，DQA 仅仅是美国政府联邦管理机构内部的管理规范，并不具有法律约束力和强制执行力，也不能作为进行任何法律诉讼的依据[59]，但可以为我国推进数据质量立法提供借鉴。笔者在 2020 年 3 月 31 日，用"数据质量"作为"标题"检索词，查询"北大法宝"(https://www.pkulaw.com/)得知，国内共出台了 32 份加强数据质量管理的部门规章，如"国家广播电影电视总局关于印发《广播影视统计数据质量管理暂行办法》的通知""国家教育委员会关于进一步提高教育统计数据质量的意见"等，但缺少专门的数据质量法律或法规。从规范数据资产、充分挖掘数据资产潜能和发挥其潜在作用以及尽可能创造最大的社会财富来看，我国应该尽早制定《数据质量管理条例》，甚至更高层次的《数据质量法》，为数据质量控制与治理提供法律支撑。

#### 3.1.2 制定《开放政府数据法》等相关法律，为科学数据质量治理提供有效的法律依据

在开放社会环境下，开放政府数据(包含政府生产或资助的科学数据，后同)成为许多国家提升政府治理能力的重要手段。2014 年 5 月 9 日，美国颁布了《2014 年数字问责与透明法》(Digital Accountability and Transparency Act of 2014，DATA)，后简称《"数据"法》。该法的主要目的是[60]：①建立政府范围内的财务数据标准，并在美国政府支出网站(usaspend.gov)或显示数据的后续系统上为纳税人和决策者提供一致、可靠和可检索的政府范围内的支出数据；②通过要求联邦机构对提交的数据的完整性和准确性负责来提高递交交给美国政府支出网站的数据质量。《"数据"法》规定了 3 项指令：①要求财政部、管理和预算办公室(OMB)为各机构提交的所有联邦支出报告创建和维护标准的数据要素和格式，并指导各机构遵循这些数据标准；②指示财政部和 OMB 以统一的开放数据集方式汇编上报的信息；③OMB 必须进行一项试点计划，以测试对接受联邦赠款和合同的人提交的报告实施数据标准的可行性。这些指令的实施有效提升了数据质量，但仍面临技术与文化方面的众多挑战，如一些机构可能认为《"数据"法》是一种官僚主义的要求，而不是一种真正的新方法；公共利益群体可能缺乏使用《"数据"法》指定数据所需的专业知识和动机等[61]。在此情况下，2019 年 1 月 14 日，美国颁布了《开放政府数据法》(Open Government Data Act)，作为《循证决策基础法》(Foundations for Evidence Based Policymaking Act)中的第二部分。《开放政府数据法》要求联邦政府机构以默认方式将政府数据公开，包括制定和维护战略信息资源管理计划(包含开放数据计划)；制定、更新和维护全面的数据清单和联邦数据目录；设立首席数据官(Chief Data Officer，CDO)和首席数据官委员会；开展报告与评估等[62]。这些措施为确保政府数据质量及其开放共享提供了法律保障。虽然政府数据开放共享已被纳入我国的大数据战略，最近新修订了《中华人民共和国政府信息公开条例》，国内出台了 50 多份地方政府促进政府数据开放共享的政策文件，如《贵阳市政府数据共享开放条例》《上海市公共数据开放暂行办

法》等,但是整体上我国政府数据开放立法还处于起步和探索阶段[63],特别是缺少国家层面的开放政府数据法,致使数据质量保证没有在国内形成统一的规范,也缺少有效的法律支撑。因此,我国应该借鉴国外成功经验,推进专门的开放政府数据立法,明确政府数据开放的范围、质量和安全要求,为政府数据质量治理提供有效的法律保障。

### 3.1.3 修订与完善现有的《科学数据管理办法》等法规或规章,为科学数据质量治理提供有效的行动指南

目前我国已经出台与科学数据管理相关的多项法规或规章(见表 2),它们虽然都提及到科学数据质量及其管理,但是除《科技基础性工作专项项目科学数据汇交管理办法(试行)》以外,其他往往在如何确保科学数据质量方面缺少实操内容。我国已经颁布《信息技术数据质量评价指标》(GB/T36344 –2018)国家标准,确立了数据质量应该满足规范性、完整性、准确性、一致性、时效性、可访问性的要求。如何通过执行该标准并由此提升科学数据质量,最有效的方法是把数据质量要求写入修订的《科学数据管理办法》,或者制定“科学数据质量管理细则”,明确利益相关者在科学数据质量治理中的权利与义务,明确科学数据质量管理与监督流程,为科学数据质量治理提供有效的行动指南。

表 2 与科学数据质量管理相关的国内法规或规章

| 法规或规章名称 | 颁布者 | 颁布时间 | 与科学数据质量管理相关的核心内容 |
|---|---|---|---|
| 《科学数据管理办法》[4] | 国务院办公厅 | 2018 年 3 月 17 日 | (1)有关科研院所、高等院校和企业等法人单位按照有关标准规范进行科学数据采集生产、加工整理和长期保存,确保数据质量;<br>(2)法人单位应建立科学数据质量控制体系,保证数据的准确性和可用性 |
| 《促进大数据发展行动纲要》[64] | 国务院 | 2015 年 8 月 31 日 | 推进数据采集、政府数据开放、数据质量等关键共性标准的制定和实施 |
| 《大数据产业发展规划(2016 –2020 年)》[65] | 工业和信息化部 | 2016 年 12 月 18 日 | (1)开展数据开放共享、产品评价、数据质量、数据安全等关键标准的试验验证和符合性检测;<br>(2)推动制定公共信息资源保护和开放的制度性文件以及政府信息资源管理办法,逐步扩大开放数据的范围,提高开放数据质量 |
| 《国家健康医疗大数据标准、安全和服务管理办法(试行)》[66] | 国家卫生健康委员会 | 2018 年 7 月 12 日 | 责任单位采集健康医疗大数据,应当严格执行国家和行业相关标准和程序,符合业务应用技术标准和管理规范,保证服务和管理对象在本单位信息系统中身份标识唯一、基本数据项一致,所采集的信息应当严格实行信息复核终审程序,做好数据质量管理 |
| 《月球与深空探测工程科学数据管理办法》[67] | 国防科工局与国家航天局 | 2016 年 9 月 12 日 | 航天器发回并经预处理的数据分为 0、1、2 三级 |
| 《科技基础性工作专项项目科学数据汇交管理办法(试行)》[68] | 科技部基础研究司和科研条件与财务司 | 2014 年 5 月 13 日 | (1)项目依托部门相关单位负责组织本部门项目的科学数据整理工作,确保数据质量;<br>(2)项目承担单位负责项目科学数据的整理和汇交,需要确保项目数据的完整性和质量;<br>(3)项目数据汇交方案内容应包括:项目基本信息、科学数据集(库)名称及主要内容、科学数据类型、科学数据格式、保密级别、保护期限、共享方式、数据质量承诺书、相关软件工具等;<br>(4)科学数据管理机构在收到项目汇交科学数据后,应在一个半月内组织完成数据测试、质量审查和验收工作 |



## 3.2 组织管理方面的治理对策

许多科学数据开放共享中的数据质量问题,如数据输入错误、数据更新错误、数据不完整、数据不一致、数据及时性差等,都与组织管理不妥有关。因此,需要从组织管理方面采取有效的治理措施,主要包括如下 3 方面:

### 3.2.1 制定科学数据质量战略,明确科学数据质量治理重点与方向

科学数据质量战略是机构对科学数据质量愿景、目标、任务以及相关的数据质量管理流程、活动与人员的规划。科学数据质量战略应该:①确立机构科学数据质量的愿景与目标,明确科学数据质量管理与质量治理的重点与方向;②确定科学数据质量的框架、维度与评价指标,制定科学数据质量测评与治理的标准;③明确科学数据质量生命周期管理,包括正规地识别、记录、检查、验证和评价科学数据质量的方法与程序;④明确科学数据质量治理的奖罚机制,打造积极高效的数据治理组织文化。一个组织或机构只有制定合理的科学数据质量战略,才能把握数据质量治理的重点与方向,助力数据质量治理的实施。

### 3.2.2 建立健全科学数据质量治理结构,明确治理主体的职责与作用

科学数据质量治理结构可以是一种四层模型,如图 2 所示:

盛小平，田婧，向桂林.科学数据开放共享中的数据质量治理研究[J].图书情报工作,2020,64(22):11－24.

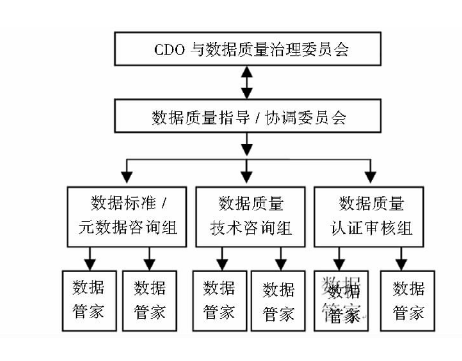


图 2　机构科学数据质量治理结构

（1）CDO 与数据质量治理委员会。在此模型中，处于治理结构最高层的是 CDO 与数据质量治理委员会。其中，CDO 主要负责[62]：①全生命周期的数据管理；②与机构内负责使用、保护、传播和生产数据的任何人员进行协调，以确保满足机构的数据需要；③管理机构的数据资产，包括数据格式的标准化、数据资产的共享与依法公开；④支持机构绩效改进人员、评价人员获取数据来履行职能；⑤在切实可行的范围内，确保机构数据符合数据管理最佳实践，最大限度地使用机构内的数据；⑥让机构雇员、公众和承包商使用公共数据资产，并鼓励采取合作方式改进数据使用；⑦识别质量改进措施需求，定期报告数据质量改进情况；⑧审查机构基础设施对数据资产易用性的影响，并与机构首席信息官协调改善这种基础设施；⑨规划与主持数据质量治理委员会工作。数据质量治理委员会指导和监督数据质量活动，主要负责：①制定机构数据质量治理政策，并提供战略指导；②审查机构数据政策，指定工作组将业务规则转换为数据规则；③批准机构数据质量政策和流程；④制定和维护数据质量标准；⑤审查数据质量实践和流程的建议；⑥签署数据质量认证和审核；⑦设置数据质量优先等级，监测并报告绩效；⑧推荐与监督 CDO。

（2）数据质量指导/协调委员会。数据质量指导/协调委员会是在数据质量治理委员会的指导下开展工作，主要负责[69]：①酌情调整数据质量治理结构中的工作组，以确保数据质量期望持续得到满足；②推荐一些合适人员加入质量治理工作组或数据质量指导委员会，协同数据质量治理委员会监督 CDO 的工作；③监督数据标准咨询组有关数据质量的工作；④建议数据质量治理委员会批准出版数据质量治理报告；⑤向数据质量治理委员会推荐数据质量标准以获得最终批准；⑥建议数据质量治理委员会批准认证和审核；⑦引导、宣传和推进已开发的数据质量实践与流程；⑧提名数据管家（data steward）；⑨参加定期会议，提供进度报告与审核状态，讨论和审查数据质量活动的总体方向；⑩参与并联系外部标准制定机构。

（3）数据标准/元数据咨询组、数据质量技术咨询组和数据质量认证审核组。数据标准/元数据咨询组负责监督各种数据标准和元数据活动，并编制和维护元数据和数据标准。具体来说，数据标准/元数据咨询组负责：①促进数据标准活动；②提供有关数据标准活动的最新报告；③管理数据标准指南使用说明文件的编制；④提供数据与元数据质量标准的培训和知识传递；⑤参与政府或行业的数据/元数据标准制定机构，执行最新的数据/元数据标准；⑥开发符合数据标准的数据质量实践，为数据标准实践提供指导[69]。

数据质量技术咨询组由负责数据质量活动技术方面的成员组成，其职责主要是：①向数据质量指导/协调委员会报告，并向委员会提供数据质量技术咨询；②更新和维护所有数据质量技术规范；③提供有关数据质量的技术和架构问题的指导；④监督数据质量技术的要求和获取过程。

数据质量认证审核组负责所有工作组的认证和审核，包括制定并发布正式的认证标准和流程，以审核数据质量政策的符合性[69]。

（4）数据管家。数据管家是工作组的成员，主要负责[69]：①收集、核对和分类数据问题，与用户和本机构员工就数据质量问题进行沟通；②管理元数据；③参与数据质量标准的制定工作；④维护数据，包括制定数据定期更新计划，保证数据资源是可用的，处理数据老化或保留问题等；⑤监督数据质量；⑥验证数据；⑦高效传播参考数据或信息；⑧管理数据源与业务规则，包括管理参考信息源，记录所有元数据、相关业务规则、数据用户与使用方式；⑨管理数据生命周期，在整个数据生命周期内解决与数据使用和可用性有关的任何问题。

通过建立上述数据质量治理结构和明确治理主体的职责，可以确保数据质量治理在机构范围内得到所有部门与员工的通力合作与广泛参与，从而确实提升科学数据质量治理水平。

### 3.2.3　制定科学数据质量治理计划，明确科学数据质量治理路径

在实施科学数据开放共享过程中，机构既要制定科学数据管理计划，也要制定科学数据质量治理计划，使两者有效融合起来。科学数据质量治理计划是对科学数据质量治理目标、任务、活动、路径与方法的规划与设计。其中，最重要的工作是评估科学数据质量现

状，量化科学数据质量的影响，确立科学数据质量治理路径，实施科学数据的清理、修正、综合、监测和报告。确立科学数据质量治理路径是实施科学数据质量治理的重要保证。一般来说，机构科学数据质量治理路径主要包括如下 10 个环节：①识别机构数据质量目标，明确机构数据质量治理领域、重点与方向；②收集、编译和分析数据质量环境信息，设计数据采集和评估计划；③评估数据质量，包括数据的准确性、完整性、一致性、及时性、可靠性、关联性、可访问性、开放性评估；④评估业务影响：使用各种技术，确定劣质数据对业务的影响，为发现根本原因、所需的数据修正提供基础；⑤确定根本原因：识别和优先考虑引发数据质量问题的真正原因，制定解决这些问题的具体建议；⑥制定改进计划：确定具体的行动建议，根据建议制定和执行改进计划；⑦防止未来的数据错误：实施处理数据质量问题根本原因的解决方案；⑧更正当前数据错误；⑨实施控制：监控、验证和维持所实施的数据质量改进；⑩沟通行动和结果：记录和交流数据质量测试结果、所做的数据质量改进以及这些改进的结果[70]。采用上述治理路径，有助于使数据质量治理理念转化为实际行动，提高数据质量治理效率。



### 3.3　技术与平台方面的治理对策

信息技术与科学数据共享平台能够显著影响科学数据质量，并为解决科学数据开放共享中的数据质量问题提供方案、模型、标准、接口等。目前我国已经建成了 8 个专门的国家科学数据共享平台和其他一系列国家科技资源共享服务平台，为科学数据开放共享奠定了坚实的设施基础。为了解决如今科学数据开放共享中的数据质量问题，仍需采取一些有效的数据治理对策。这主要包括：

#### 3.3.1　把数据剖析嵌入科学数据质量管理流程，增强数据质量治理效果

数据剖析（data profiling）是审查源数据与理解数据结构、内容和相互关系以及识别数据项目潜力的过程[71]，也是对数据库元数据（包括列和记录等数据源的当前状态）进行分析，并检查合理的数据位置、数据结构和数据值的过程[72]。这些元数据可以从简单的统计数据，如列中的空值和不同值的数量以及列的数据类型或数据值的最常见模式，到复杂的值间和列间的依赖关系[73]。数据剖析包括 5 方面工作任务[74]：①元数据分析：发掘元数据信息，如数据结构、数据创建者、创建时间等；②数据表示分析：查找数据模式，包括文本模式、时间模式和数字模式，如地址模式、日期模式和电话模式；③数据内容分析：审查数据基础信息，包括数据的准确性、及时性、完整性等；④数据集分析：分析来自数据集中的数据，例如统计、分布、基数、频率、最大值或最小值、平均值、冗余等；⑤数据逻辑规则分析：根据业务逻辑规则或数据的逻辑含义、业务规则和功能依赖关系审查数据。通过执行上述任务，数据剖析可以验证用户结构化数据、半结构化数据和非结构化数据，收集数据结构、数据模式、统计信息、分发消息，审查数据治理、数据管理、数据迁移和数据质量控制的数据属性。由此看来，数据剖析具有评估数据质量和改进数据质量的功能，可以发现数据质量问题，提高数据源的可靠性和完整性。因此，需要把数据剖析嵌入数据质量管理流程。这种科学数据质量管理流程主要包括如下 8 个步骤[72]：①收集与分析数据和元数据。需要收集科学数据与科学数据库的物理元数据。其中，物理元数据必须包括表和列的名称、数据类型、域信息、约束、实体关系、数据库的代码定义，并将物理元数据与数据产品如表列名称、数据类型等进行比较分析，若发现无效点，还需验证它们。②选择数据剖析来源和类型。对科学数据来源、表格和分析类型进行分析和检查。③对元数据和数据源进行数据剖析。分析所选的表、列和数据源。通过对数据状态的分析，可以发现遗漏值、无效值、非唯一值、数据和结构完整性的破坏。如果发现错误或无效数据，则需要进行验证。④审查和报告数据剖析。与科学数据业务主管综合数据剖析和审查结果。业务主管需要确认使用数据剖析得到的无效数据和错误状态。通过与业务主管的讨论，制定修改无效数据和错误状态的业务规则。⑤数据抽取。提取无效或错误数据，并转换或清除这些数据，如转换数据类型、代码定义、数据格式等，或清除表名、列名、无效数据、空值、键数据等。⑥数据转换。根据业务规则或质量管理程序转换数据。⑦数据清理。业务主管动手清理数据，并对无效数据进行修改以保证科学数据的可靠性。⑧数据加载。将经过转换和清理的数据加载到数据库中，以便在机构内部或跨机构之间迁移、交换与共享科学数据。这种添加了数据剖析技术的数据质量管理流程，可以发现容易出错的地方，将异常数据与已检查的元数据和数据源区分开来，清理与修正错误数据，实现数据质量治理的目的。

#### 3.3.2　实施科学数据质量审计，明确数据质量治理的重点领域

科学数据质量审计是基于科学数据质量标准，对科学数据的各种属性是否达到标准要求以及存在哪些

盛小平，田婧，向桂林. 科学数据开放共享中的数据质量治理研究[J]. 图书情报工作,2020,64(22):11－24.

质量问题的检查与评估。从理论上讲,科学数据的准确性、完整性、一致性、及时性、可靠性、关联性、可访问性、开放性越高越好,但在开放共享实践中确保高品质科学数据并非易事。这时,建立科学数据质量清单,实施科学数据质量审计,是发现科学数据开放共享中的数据质量问题的最有效办法。这种科学数据质量清单需要覆盖开放共享科学数据的 8 种属性见表 3,且按 5 分制(0,1,2,3,4)来评估科学数据属性的得分,其中,4 分表示得到非常好的处理(即具有最高的质量);3 分表示虽得到处理,但仍需改进;2 分表示得到部分处理,需要很大的改进;1 分表示根本没有处理;0 分表示完全不适用。数据专业人员可以利用这种数据质量清单,开展科学数据质量审计,由此发现所在机构的科学数据存在哪些质量缺陷以及需要进行何种层次的质量改进,从而明确科学数据质量治理的重点领域与核心工作,更好地促进科学数据质量治理的实施。

**表 3 开放共享的科学数据质量清单**

| 准确性 | 评分 | 一致性 | 评分 | 开放可访问性 | 评分 |
|---|---|---|---|---|---|
| 数据内容准确 | | 数据内容一致 | | 数据是开放共享的 | |
| 数据类型准确 | | 数据结构一致 | | 可以访问原始数据 | |
| 数据语法准确 | | 数据类型一致 | | 数据可供检索 | |
| 数据引用准确 | | 数据标准一致 | | 机器可读格式 | |
| 数据无污染 | | 数据格式一致 | | 数据软件兼容 | |
| 数据无碎片化 | | 数据注释一致 | | 数据可进行开放验证 | |
| 可靠性 | 评分 | 完整性 | 评分 | 关联性 | 评分 |
| 数据来源可靠 | | 数据质量信息完整 | | 满足研究的需要 | |
| 数据内容可靠 | | 数据内容完整 | | 数据已链接 | |
| 数据位置明确 | | 数据值完整 | | 数据链接正确 | |
| 数据是安全的 | | 数据引用完整 | | 数据链接可靠 | |
| 数据是可验证的 | | 元数据完整 | | | |
| 及时性 | 评分 | | | | |
| 数据是最新可用的 | | | | | |
| 数据收集后尽快报告 | | | | | |
| 数据及时更新 | | | | | |

#### 3.3.3 构建关联开放数据,增强科学数据的关联性、开放可访问性

不同于一般的开放数据,关联开放数据需满足 5 种要求[75]:①在网络上可用,提供的任何格式数据都是开放许可的;②发布机器可读的结构化数据;③提供非专有格式;④利用 W3C 中的开放标准,如资源描述框架(RDF)和查询语言 Sparql 等通过统一资源标识符(URI)来识别事物;⑤将数据链接到其他数据集以提供上下文关系。

由于具有上述特性,关联开放数据成为实现开放数据资源整合的新技术与新方法,能够摆脱现有 Web 网络信息的粗粒度与语义缺失的现象[76]。它通过发布和链接结构化的数据使分散异构的数据实现语义关联和集成,可以提高开放科学数据质量。例如,Springer Nature 推出了一个学术领域的关联开放数据平台——Springer Nature SciGraph。它集成 Springer Nature 及其学术领域合作伙伴的数据资源,如相关资助机构、研究机构、科研项目、会议、出版物等各个研究领域的信息,同时不断从各种数字资源如数字期刊、文章、图书和章节、专利、会议、引用和参考链接网络等获取更多的元数据,能够提供更加丰富的语义描述,让更多来自可靠来源的高质量信息可被发现和利用,从而可以帮助科研共同体充分利用开放科学数据,促进学术交流与创新[77]。因此,通过构建关联开放数据与搭建关联开放数据网络,提高语义网中的数据集质量、链接质量和模式质量,就可获得科学数据质量保证[78],增强科学数据的关联性、开放性、可访问性和互操作性。

### 3.4 利益相关者方面的治理对策

从利益相关者角度来看,科学数据质量治理对策主要包括如下两方面:

#### 3.4.1 明确不同利益相关者的科学数据质量治理职责与作用

目前缺少相关法规或规章对利益相关者(包括政府、研究人员、研究机构、数据中心、图书情报机构、资助机构、出版社、数据专业人员等)在科学数据开放共

享中的数据质量治理职责进行有效界定,使数据质量治理收效甚微。因此,十分有必要明确不同利益相关者的数据质量治理职责与作用。

(1)政府。笔者认为,政府是实施科学数据质量治理不可或缺的重要因素。一方面,政府是科学数据管理政策与法规的制定者,可以为科学数据质量治理创建良好的法制环境与政策环境;另一方面,一些政府机构也是科学数据的生产者、出版者、传播者和管理者,政府机构提供的科学数据质量的高低直接影响公众对这些科学数据利用效果的好坏,甚至影响国家大数据战略的实施。因此,应该充分发挥政府在制定科学数据管理与质量治理政策与法规上的主导作用,通过建立健全我国科学数据管理与质量治理政策与法规来保证科学数据质量治理的有效实施。

(2)研究人员和研究机构。研究人员既是科学数据的主要生产者与使用者,也是科学数据质量治理的中坚力量。一方面,研究人员应该克服科学数据开放共享的认知障碍,打破"不愿开放和不敢开放"的禁锢,积极创建、共享与利用有价值的科学数据;另一方面,研究人员应该确保开放科学数据的准确性、完整性、一致性、及时性、可靠性、关联性、可访问性与开放性,提供科学数据质量保证,使自己成为科学数据质量治理的践行者与中坚力量。

研究机构作为科学数据的生产者、管理者、传播者与利用者,在数据质量治理中具有独特的作用。研究机构不仅要为科学数据开放共享创造有利的内部环境,开发支撑科学数据开放共享的基础设施,提供机构开放科学数据的长期保存与访问[79],而且需要制定本机构科学数据开放共享政策与数据质量标准,建立科学数据质量治理结构与规则,明确机构科学数据质量治理路径,增强机构科学数据质量治理的内生力量。在此方面,《中国科学院科学数据管理与开放共享办法(试行)》起到了示范作用。该办法中的许多条款明确了研究机构的科学数据质量治理责任与作用,比如:第7条规定中国科学院网络安全和信息化领导小组办公室要负责全院科学数据管理与开放共享的标准化工作;第9条规定院属法人单位要建立健全科学数据管理与开放共享制度和科学数据质量控制体系;第19条强调科学数据应按照分等级、可发现、可访问、可重用的原则,适时向院内外用户开放共享;第23条规定中国科学院科学数据中心要开展科学数据加工与质量控制工作,形成分级分类开放共享的目录清单;第30条规定对于伪造、篡改、剽窃、抄袭、重复出版科学数据等严重科研不端行为,将按院有关制度进行学术调查并给予相应学术处理[80]。这些条款奠定了中国科学院科学数据质量治理的基础,值得其他机构借鉴。

(3)数据中心与图书情报机构。数据中心与图书情报机构作为科学数据的主要组织者、发布者、传播者、管理者与服务提供者,在科学数据质量治理中具有重要的地位。他们应该制定科学数据开放共享政策,明确科学数据质量技术标准,建立健全科学数据质量管理生命周期与质量管理流程,动态评估与监管本机构组织、收藏与发布的科学数据质量,履行科学数据质量分析师、评价者与监管者或控制者角色,为科学数据质量治理提供支撑与保障。

(4)资助机构。资助机构在科学数据质量治理方面具有激励与引导作用,不仅要求申请者在申请资助项目时需要提交其科学数据管理计划,而且明确规定科学数据的质量要求与问责,并把这些内容写入资助协议中,这样既能约束申请者恪守数据质量关,也能发挥资助者的监督作用。例如,美国国家科学基金会规定:申请人在提交项目申请书必须包含一份"数据管理计划",详细说明[81]:①项目过程中需要制作的数据类型、样本、实物、软件、课程资料和其他资料;②用于数据和元数据格式与内容的标准(如果现有标准缺失或被认为不足,应将其与任何建议的解决方案或补救措施一起记录在案);③访问和共享政策,包括适当保护隐私、机密性、安全性、知识产权、其他权利或要求的条款;④关于重复使用、重新分配和衍生品制作的政策和规定;⑤数据、样本和其他研究产品的存档与保存访问的计划。这种资助政策有助于实现科学数据质量治理。

(5)出版社。出版社是科学数据的主要发布者与传播者。出版社通过建立科学数据开放出版规则,规范科学数据出版质量要求,严格控制科学数据的出版发行,使得只有符合科学数据质量标准的数据才能发布出来,从而发挥其作为科学数据质量监管者、守门人的作用。这也是科学数据质量治理过程中不可或缺的一个关键环节。特别是近年来,部分国际知名出版商纷纷推出数据期刊,如 Springer-Nature 创办的 *Scientific Data*、Wiley- Blackwell 创办的 *Geoscience Data Journal*、*Elsevier* 创办的 *Data in Brief* 等,均要求作者对发表的数据集进行详细描述,说明数据来源、处理过程、使用的软件和数据文件类型等,并采用严格的同行评审机制和数据引用政策,确保论文中的数据质量达到较高水平。

（6）数据专业人员。数据专业人员主要包括数据开发者、数据研究者、数据创建者、数据管理者、数据服务提供者，常见的称谓有数据科学家、首席数据官、数据分析师、数据库架构师、数据可视化专家、数据质量经理、数据馆员、数据管家、数据安全官[82]等。各种机构应该设置数据专业人员职位，明确其在科学数据管理与质量治理中的职责，包括承担科学数据质量检查、审核、评估、修正、报告的具体工作任务，并赋予其管理、控制机构科学数据质量的权力，从而确保开放共享的科学数据具有较高的质量。

### 3.4.2 建立利益相关者科学数据质量协同治理机制

正因为众多利益相关者在科学数据开放共享中的数据质量治理中有不同的职责与作用，所以需要建立利益相关者科学数据质量协同治理机制以便更高效实施数据质量治理。科学数据质量协同治理机制是指多元合法治理主体（如政府部门、各种机构、公众等）基于法律法规和其他行为规范，跨越组织边界，通过相互配合与协同来解决科学数据质量问题和获得高质量数据的作用机理与运行方式。协同治理的本质要义在于打破不同治理主体之间的层层障碍，利用质量治理社会网络中节点之间错综复杂的社会关系，协同处理科学数据质量问题，并为科学数据质量提供保障。实施这种协同治理机制的关键在于：①通过上级政府部门或主管机构、资助机构对科学数据质量治理的顶层设计与制度安排，包括制定相关科学数据质量治理政策、明确不同利益相关者在科学数据质量治理中的职责等，为科学数据质量协同治理提供政策保障；②通过利用治理主体（即质量治理社会网络中的节点）之间的各种社会关系（如领导、资助、管理、监督、协同、合作等），发挥治理主体联动作用，协同解决单个治理主体无法处理的科学数据质量问题；③通过建立科学数据质量治理的利益驱动机制，包括建立科学数据质量治理的奖惩机制、信誉机制，引导和促进利益相关者积极参与科学数据质量的协同治理。

## 4 结语

数据质量是科学数据开放共享必须重视的一个关键问题。一方面，科学数据开放共享是否取得成功在很大程度上依赖于高质量的科学数据；另一方面，目前科学数据开放共享遇到了一些数据质量问题，如科学数据的准确性、完整性、一致性、及时性、可靠性、关联性、开放可访问性等。产生科学数据质量问题的根本原因主要来自于政策法规、组织管理、技术与平台、利益相关者等方面。针对科学数据开放共享中的各种数据质量问题，可以从其诱因入手，制定有效的数据质量治理对策，包括：①在政策法规方面，制定《数据质量法》或《数据质量管理条例》，为科学数据质量治理提供专门的法律支撑；制定《开放政府数据法》等相关法律，为科学数据质量治理提供有效的法律依据；修订与完善现有的《科学数据管理办法》等法规或规章，为科学数据质量治理提供有效的行动指南。②在组织管理方面，制定科学数据质量战略，明确科学数据质量治理重点与方向；建立健全科学数据质量治理结构，明确治理主体的职责与作用；制定科学数据质量治理计划，明确科学数据质量治理路径。③在技术与平台方面，把数据剖析嵌入科学数据质量管理流程，增强数据质量治理效果；实施科学数据质量审计，明确数据质量治理的重点领域；构建关联开放数据，增强科学数据的关联性、开放性与可访问性。④在利益相关者方面，明确不同利益相关者的科学数据质量治理职责与作用，并建立利益相关者科学数据质量协同治理机制。这些数据质量治理措施将有效解决科学数据开放共享中的数据质量问题，有助于进一步推动科学数据的开放共享和更大范围内的开放研究与开放创新。

**参考文献：**

[1] LAMPOLTSHAMMER T J, SCHOLZ J. Open data as social capital in a digital society [M]//KAPFERER E, GSTACH I, KOCH A, et al. Rethinking social capital: global contributions from theory and practice. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017:137－150.

[2] 张坦，黄伟，石勇. ISO 8000（大）数据质量标准及应用[J]. 大数据,2017(1):3－11.

[3] G8. G8 open data charter [EB/OL]. [2020－06－06]. http://opendatacharter. net/wp-content/uploads/2015/10/opendatacharter-charter_F. pdf.

[4] 国务院办公厅. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2020－06－06]. http://www. most. gov. cn/mostinfo/xinxifenlei/fgzc/gfxwj/gfxwj2018/201804/t20180404_139023. htm.

[5] 王志强，杨青海. 科学数据质量及其标准化研究[J]. 标准科学,2019(3):25－30.

[6] MADINICK S, WANG R, XIAN X. The design and implementation of a corporate householding knowledge processor to improve data quality [J]. Journal of management information systems, 2004(1):41－49.

[7] ZUIDERWIJK A, JANSSEN M, CHOENNI S, et al. Socio-technical impediments of open data [J]. Electronic journal of e-government, 2012, 10(2):156－172.

[8] JANSSEN M, CHARALABIDIS Y, ZUIDERWIJK A. Benefits, adoption barriers and myths of open data and open government[J].

Information systems management,2012,29(4):258 - 268.
[9] GEIGER J G. Data quality management: the most critical initiative you can implement [EB/OL]. [2020 - 06 - 06]. https://support. sas. com/resources/papers/proceedings/proceedings/sugi29/098-29. pdf.
[10] MORBEY G. Data quality for decision makers[M]. 2nd ed. Wiesbaden: Springer Gabler, 2013.
[11] 刘冰,庞琳. 国内外大数据质量研究述评[J]. 情报学报,2019,38(2):217 - 226.
[12] KULIKOWSKI J L. Data quality assessment: problems and methods [J]. International journal of organizational and collective intelligence, 2014, 4(1):24 - 36.
[13] ISO/IEC 25012[EB/OL]. [2020 - 06 - 06]. https://iso25000.com/index. php/en/iso-25000-standards/iso-25012.
[14] LEE W Y,PIPINO L L,FUNK J D, et al. Journey to data quality [M]. Cambridge: The MIT Press, 2006.
[15] HAUG A, ARLBJØRN J S. Barriers to master data quality [J]. Journal of enterprise information management, 2011, 24(3):288 - 303.
[16] OLSON J E. Data quality: the accuracy dimension [M]. San Francisco: Morgan Kaufmann Publishers, 2003.
[17] WAND Y, WANG R. Anchoring data quality dimensions in ontological foundations [J]. Communications of the ACM, 1996, 39 (11):86 - 95.
[18] BATINI C, SCANNAPIECO M. Data and information quality: dimensions, principles and techniques [M]. Cham: Springer International Publishing AG, 2016.
[19] LI X, ZHAI J, ZHENG G, et al. Quality assessment for open government data in China [EB/OL]. [2020 - 06 - 06]. https://dl.acm. org/doi/pdf/10. 1145/3285957. 3285962.
[20] KIM W, CHOI B J, HONG E, et al. A taxonomy of dirty data [J]. Data mining and knowledge discovery, 2003, 7(1):81 - 99.
[21] 李晓彤,翟军,郑贵福. 我国地方政府开放数据的数据质量评价研究——以北京、广州和哈尔滨为例[J]. 情报杂志, 2018,37(6):141 - 145.
[22] CSÁKI C. Towards open data quality improvements based on root cause analysis of quality issues[J]. Lecture notes in computer science, 2018,11020:208 - 220.
[23] TAYI G K, BALLOU D P. Examining data quality [J]. Communications of the ACM, 1998, 41(2): 54 - 57.
[24] 温亮明,张丽丽,黎建辉. 大数据时代科学数据共享伦理问题研究[J]. 情报资料工作,2019,40(2):38 - 44.
[25] STAGARS M. Open data in Southeast Asia[M]. Singapore: Palgrave Macmillan, 2016.
[26] 夏姚璜,邢文明. 开放政府数据评估框架下的数据质量调查与启示[J]. 情报理论与实践, 2019, 42(8):44 - 49,66.
[27] DAMA International. Data management body of knowledge [M]. 2nd ed. Basking Ridge: Technics Publications, 2017.
[28] ZAVERI A, KONTOKOSTAS D, SHERIF M A, et al. User-driven quality evaluation of dbpedia [EB/OL]. [2020 - 06 - 06]. http://svn. aksw. org/papers/2013/ISemantics_DBpediaDQ/public. pdf.
[29] BEHKAMAL B, KAHANI M, BAGHERI E, et al. A metrics-driven approach for quality assessment of linked open data[J]. Journal of theoretical and applied electronic commerce research, 2014,9 (2):64 - 79.
[30] 王春山. 数据质量管理在银行信用卡数据管理中的应用[D]. 广州:华南理工大学,2005.
[31] LARANJEIRO N, SOYDEMIR S N, Bernardino J. A survey on data quality: classifying poor data [EB/OL]. [2020 - 06 - 06]. https://ieeexplore. ieee. org/stamp/stamp. jsp? tp = &arnumber = 7371861.
[32] 温芳芳. 我国政府数据开放的政策体系构建研究[D]. 武汉:武汉大学,2019.
[33] SAMUEL-ROSA A, DALMOLIN R S D, MOURA-BUENO J M, et al. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it[EB/OL]. [2020 - 06 - 06]. http://www. revistas. usp. br/sa/article/view/160727/154973.
[34] SCHMIDT B, GEMEINHOLZER B, TRELOAR A. Open data in global environmental research: the Belmont Forum's open data survey[EB/OL]. [2020 - 06 - 06]. https://journals. plos. org/plosone/article? id = 10. 1371/journal. pone. 0146695.
[35] LEI Y, NIKOLOV A, UREN V, et al. Detecting quality problems in semantic metadata without the presence of a gold standard[EB/OL]. [2020 - 06 - 06]. http://ceur-ws. org/Vol-329/paper06. pdf.
[36] 郭路生,刘春年. 大数据时代应急数据质量治理研究[J]. 情报理论与实践,2016,39(11):101 - 105.
[37] 夏姚璜,邢文明. 开放政府数据评估框架下的数据质量调查与启示[J]. 情报理论与实践, 2019, 42(8):44 - 49,66.
[38] CONRADIE P, CHOENNI S. On the barriers for local government releasing open data[J]. Government information quarterly, 2014, 31(S1): 10 - 17.
[39] 盛小平,吴红,胡冰洁. 科学数据开放共享障碍的实证分析[J]. 图书情报工作,2019,63(17): 23 - 30.
[40] CAI L, ZHU Y. The challenges of data quality and data quality assessment in the big data era [EB/OL]. [2020 - 06 - 06]. https://datascience. codata. org/articles/10. 5334/dsj-2015-002/.
[41] NI K, CHU H, ZENG L, et al. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study [EB/OL]. [2020 - 06 - 06]. https://www. ncbi. nlm. nih. gov/pmc/articles/PMC6609143/pdf/bmjopen-2019-029314. pdf.
[42] 闫桂勋,刘蓓,程浩,等. 数据共享安全框架研究[J]. 信息安全研究,2019,5(4):309 - 317.
[43] LEVIN N, LEONELLI S, WECKOWSKA D,et al. How do scientists define openness? exploring the relationship between open sci-

ence policies and research practice[J]. Bulletin of science, technology & society, 2016,36(2):128－141.

[44] 洪学海,王志强,杨青海.面向共享的政府大数据质量标准化问题研究[J].大数据,2017,3(3): 44－52.

[45] KUULA A, BORG S. Open access to and reuse of research data - the state of the art in Finland[M]. Tampere: Finnish Social Science Data Archive,2008.

[46] PANHUIS W G V, PAUL P,EMERSON C, et al. A systematic review of barriers to data sharing in public health[EB/OL].[2020－06－06]. https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-1144.

[47] 林焱.我国政府数据开放的元数据管理研究[D].武汉:武汉大学,2018:57.

[48] GADE S. LinkWiper-a system for data quality in linked open data[D]. Dearborn:University of Michigan-Dearborn, 2016.

[49] WRIGHL S, GENAL O. Data quality assessment[M]. Bradley Beach: Technics Publications, LLC, 2007.

[50] ZUIDERWIJK A,JANSSEN M. Open data policies, their implementation and impact: a framework for comparison[J]. Government information quarterly,2014,31(1):17－29.

[51] 王娟.国内外政府开放数据质量研究述评[J].图书馆理论与实践,2019(12):27－31.

[52] TSOUKALA V, ANGELAKI M, KALAITZI V, et al. Policy recommendations for open access to research data in Europe[EB/OL].[2020－06－06]. https://ec.europa.eu/newsroom/dae/document.cfm? doc_id=9958.

[53] VOULGARIS Z. Data scientist: the definitive guide to becoming a data scientist[M]. Basking Ridge:Technics Publications, 2014.

[54] OTTO B. On the evolution of data governance in firms: the case of Johnson & Johnson consumer products North America[M]// SADIQ S. Handbook of data quality: research and practice. Berlin: Springer-Verlag,2013:93－118.

[55] ZHANG J. Operationalizing data quality through data governance [M]//BHANSALI N. Data governance: creating value from information assets. Boca Raton: CRC Press,2014:65－92.

[56] IBM. What is data governance? [EB/OL].[2020－06－06]. https://www.ibm.com/analytics/data-governance.

[57] KOLTAY T. Quality of open research data: values, convergences and governance[EB/OL]. [2020－06－06]. https://www.mdpi.com/2078-2489/11/4/175/pdf.

[58] FinDLaw Attorney Writers. Federal agencies subject to data quality act[EB/OL]. [2020－06－06]. https://corporate.findlaw.com/law-library/federal-agencies-subject-to-data-quality-act.html.

[59] 宋立荣,彭洁.美国政府"信息质量法"的介绍及其启示[J].情报杂志,2012,31(2):12－18.

[60] Digital accountability and transparency act of 2014[EB/OL]. [2020－06－06]. https://www.govinfo.gov/content/pkg/PLAW-113publ101/pdf/PLAW-113publ101.pdf.

[61] The Data Foundation. Data act 2022: changing technology, changing culture[EB/OL]. [2020－06－06]. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/us-ps-data-act-2022.pdf.

[62] Foundations for evidence-based policymaking act of 2018[EB/OL].[2020－06－06]. https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf.

[63] 翟军,李昊然,孙小荃,等.美国《开放政府数据法》及实施研究[EB/OL].[2020－06－06]. http://kns.cnki.net/kcms/detail/11.1762.G3.20200317.1150.002.html.

[64] 国务院.国务院关于印发促进大数据发展行动纲要的通知[EB/OL].[2020－06－06]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[65] 工业和信息化部.工业和信息化部关于印发大数据产业发展规划(2016－2020年)的通知[EB/OL].[2020－06－06]. http://www.miit.gov.cn/n1146285/n1146352/n3054355/n3057656/n5340632/c5465614/part/5465622.doc.

[66] 国家卫生健康委员会.关于印发国家健康医疗大数据标准、安全和服务管理办法(试行)的通知[EB/OL].[2020－06－06]. http://www.cac.gov.cn/2018-09/15/c_1123432498.htm.

[67] 国防科工局,国家航天局.国防科工局 国家航天局关于印发《月球与深空探测工程科学数据管理办法》的通知[EB/OL].[2020－06－06]. http://www.sastind.gov.cn/n4235/c6807016/content.html.

[68] 科技部基础研究司.科技基础性工作专项项目科学数据汇交管理办法(试行)[EB/OL]. [2020－06－06]. http://www.most.gov.cn/tztg/201406/W020140625319357180895.doc.

[69] LOSHIN D. The practitioner's guide to data quality improvement [M]. Burlington: Morgan Kaufmann, 2011:120－121.

[70] MCGILVRAY D. Executing data quality projects: ten steps to quality data and trusted information[M]. San Francisco: Morgan Kaufmann, 2008.

[71] What is data profiling? process, best practices and tools [EB/OL]. [2020－06－06]. https://panoply.io/analytics-stack-guide/data-profiling-best-practices/.

[72] KOOK Y, LEE J, PARK M, et al. Data quality management based on data profiling in e-government environments[C]//KIM T, ADELI H, ROBLES R J, et al. Advanced communication and networking. Berlin:Springer-Verlag,2011.

[73] ABEDJAN Z. Data profiling[M]//SAKR S, ZOMAYA A Y. Encyclopedia of big data technologies[M]. Basel: Springer Nature Switzerland AG, 2019:563－568.

[74] DAI W,WARDLAW I,CUI Y,et al. Data profiling technology of data governance regarding big data: review and rethinking[C]// DAI W,WARDLAW I,CUI Y,et al. Information technology: new generations. Berlin: Springer-Verlag, 2016:439－450.

[75] BERNERS-LEE T. Linked data[EB/OL].[2020－06－06]. http://www.w3.org/DesignIssues/LinkedData.html.

[76] BAUER F, KALTENBÖCK M. Linked open data: the essentials [EB/OL].[2020－06－06]. https://www.reeep.org/LOD-the-

Essentials. pdf.
[77] Springer Nature. SN SciGraph: a linked open data platform for the scholarly domain [EB/OL]. [2020 - 06 - 06]. https://www.springernature.com/gp/researchers/scigraph.
[78] HADHIATMA A. Improving data quality in the linked open data: a survey[EB/OL]. [2020 - 06 - 06]. https://iopscience.iop.org/article/10.1088/1742-6596/978/1/012026/pdf.
[79] 盛小平,王毅. 利益相关者在科学数据开放共享中的责任与作用[J]. 图书情报工作,2019, 63(17):31 - 39.
[80] 中国科学院. 中国科学院关于印发《中国科学院科学数据管理与开放共享办法(试行)》的通知[EB/OL]. [2020 - 06 - 06]. http://www.go.cas.cn/gzzd/xxhgz/201911/t20191101_4722182.html.
[81] NSF. Proposal & award policies & procedures guide (PAPPG) [EB/OL]. [2020 - 06 - 06]. https://www.nsf.gov/pubs/policydocs/pappg20_1/nsf20_1.pdf.
[82] FOSTER J, MCLEOD J, NOLIN J, et al. Data work in context: value, risks, and governance [J]. Journal of the association for information science and technology, 2018, 69(12):1414 - 1427.

**作者贡献说明:**

盛小平:论文撰写与修订;
田婧:参与论文初稿写作;
向桂林:参与论文框架设计与修订。

## Research on Data Quality Governance in Open Sharing of Scientific Data

Sheng Xiaoping[1] Tian Jing[1] Xiang Guilin[2]
[1] School of Library, Information and Archives, Shanghai University, Shanghai 200444
[2] Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101

Abstract: [**Purpose/significance**] In order to promote the effective implementation of open sharing of scientific data, this paper explores the data quality problems in open sharing of scientific data and its governance countermeasures. [**Method/process**] By means of normative analysis and causal analysis, this paper analyzed the data quality problems in open sharing of scientific data and the root causes of the problems, then constructd the governance model of open sharing of scientific data, finally proposed four types of governance countermeasures from the perspective of inducements. [**Result/conclusion**] The problems of data quality in open sharing of scientific data involve the accuracy, completeness, consistency, timeliness, reliability, relevance and open accessibility of scientific data. In order to solve the problems of scientific data quality and further promote the implementation of open sharing of scientific data, countermeasures for scientific data quality governance can be formulated from four aspects of policies and regulations, organizational managements, technologies and platforms, and stakeholders.

Keywords: scientific data open sharing data quality quality governance governance countermeasure